Rethinking Language Models Within the Framework of Dynamic Bayesian Networks

Murat Deviren, Khalid Daoudi, and Kamel Smaïli

INRIA-LORIA, Parole team, 54602 Villers les Nancy, France
{daoudi, deviren, smaili}@loria.fr

Abstract. We present a new approach for language modeling based on dynamic Bayesian networks. The philosophy behind this architecture is to learn from data the appropriate relations of dependency between the linguistic variables used in language modeling process. It is an original and coherent framework that processes words and classes in the same model. This approach leads to new data-driven language models capable of outperforming classical ones, sometimes with lower computational complexity. We present experiments on a small and medium corpora. The results show that this new technique is very promising and deserves further investigations.

1 Introduction

A statistical speech recognition system estimates the most probable sequence of linguistic units (i.e. words, syllables, phonemes, etc.) given acoustic observations. The Bayesian formulation of the problem allows a factorization over the acoustic and linguistic components: $\hat{W} = \arg\max_W P(O|W)P(W)$, where O denotes acoustic observations and W denotes the underlying sequence of linguistic units. In this formulation, the language model, $P(W)$, encodes the a priori linguistic information, i.e. syntactic, lexical and/or morphologic properties of language. The specification of a language model involves the definition of implicit and/or explicit variables of language. For example n-gram models use *word* as the only variable in language whereas syntactic n-class models use both *word* and *syntactic classes* [1,2]. The dynamics of language is derived by these variables and their interactions through time. Each variable interacts with a certain number of factors that constitute its context. In probabilistic terms the context of a linguistic unit is defined with conditional independence properties. Given its context each linguistic unit is assumed to be independent of other linguistic events. For example classical n-gram models make the assumption that a word is independent of all preceding words given the most recent $n-1$.

On the other hand, conditional independence is the core property of *dynamic Bayesian networks* (DBNs), it is indeed the exploitation of this property that leads to efficient and generic inference algorithms [3]. Moreover, as it will become clear in the following sections, n-gram and n-class models (and other language models) are very particular instances of DBNs. Thus, it is a natural idea to rethink language models within the general framework of DBNs and seek potential benefits from this rethinking.

It is our purpose in this paper to use the DBNs framework in order to achieve a better exploitation of each linguistic unit considered in modeling. We develop a unifying

B. Kégl and G. Lapalme (Eds.): AI 2005, LNAI 3501, pp. 432–437, 2005.

approach that processes each of these units in a unique model and construct new data-driven language models with improved performances. The principle of our approach is to construct DBNs in which a variable (word, class or any other linguistic unit) may depend on a set of context variables. These dependences between linguistic units can be determined automatically or manually. Of course our ultimate goal is to propose an automatic scheme to learn the optimal DBN structure from a training corpus. However, in order to investigate the feasibility of our approach, we start by analyzing DBN models for which the graphical structure is specified manually.

2 A Brief Overview of Dynamic Bayesian Networks

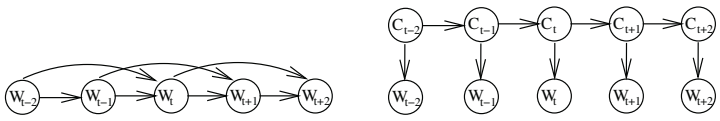
Dynamic Bayesian networks (DBNs) are generalization of (static) Bayesian networks (BNs) to dynamic processes. The Bayesian networks formalism consists of associating a directed acyclic graph to the joint probability distribution (JPD) $P(X)$ of a set of random variables $X = \{X_1, ..., X_n\}$. The nodes of this graph represent the random variables, while the arrows encode the conditional independences (CI) which (are supposed to) exist in the JPD. A DBN encodes the temporal dynamics of a time evolving set $X[t] = \{X_1[t], \ldots, X_n[t]\}$ of variables. The JPD of $\mathbf{X}_1^T = \{X[1], \ldots, X[T]\}$ is factorized as:

$$P(X[1], \ldots, X[T]) = \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i[t] | \Pi_{it}) \tag{1}$$

where Π_{it} denotes the parents of $X_i[t]$. In the BNs literature, DBNs are defined using the assumption that $X[t]$ is Markovian [4]. In this paper, we relax this hypothesis to allow non-Markov processes (see [5] for details).

From this perspective, it is obvious that classical language models can be represented as DBNs. Indeed, n-gram models assume that the probability of a word sequence is factorized over the conditional probabilities of each word in the sequence given its recent history of $n-1$ words. That is, if W is the word vocabulary and $w_1^T = w_1...w_T \in W^T$ is a word sequence, one assumes that: $P(w_1^T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_{t-n+1})$

Thus, if W_t is a discrete random variable taking its values in W for every t, n-grams can be represented as the DBN shown in Fig. 1-(a) (for $n = 3$, i.e., tri-gram) which is a Markov chain of order n. Class-based approaches represent the history on word classes rather than words. That is, if $C = \{l_1, ..., l_m\}$ is the set of class labels and $c_1^T = c_1...c_T \in C^T$ is an observed class sequence, one assumes that:



(a) tri-gram (b) bi-class

Fig. 1. Tri-gram and bi-class models

$P(w_1^T, c_1^T) = \prod_{t=1}^{T} P(w_t|c_t)P(c_t|c_{t-1}, \ldots, c_{t-n+1})$. Thus, if C_t is a discrete random variable taking its values in C for every t, n-class models can be represented as the DBN shown in Fig. 1-(b) (for $n = 2$, i.e., bi-class).

3 Language Modeling with DBNs

n-gram and n-class models are the most commonly used language models in state-of-the-art speech recognition systems. In practice they are merged together either using linear combination or an integration of their respective characteristics in a single architecture using maximum entropy techniques [6]. This approach yields quite interesting results, however if we want to better exploit the lexical and syntactic information, a solution would be to consider them in a unique model that is trained within a single procedure.

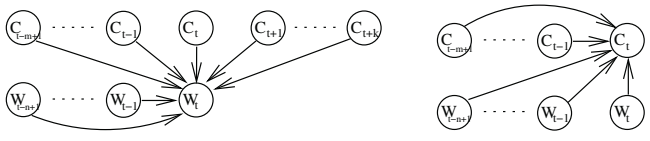
The DBN formalism provides a theoretical and computational framework to achieve this goal. Our principle idea is to impose no *a priori* hypothesis on the way a language should be represented but to consider all available data (words, classes, ...) as observations of the dynamic system $\{W_t, C_t\}$. Our goal then is to find the model that has the best description (in terms of perplexity) of these observations. In this way, we let data dictate what influences the pronunciation of a word. In Bayesian networks terminology this is the *structure learning* problem: find the graph structure (and its numerical parameterization) that explains the data at ''best".

In order to define a set of DBN structures plausible for language modeling, we need to specify conditional independence (CI) assertions that are linguistically informative and easy to interpret. We also want n-gram and n-class models to be included in this set in order to be able to exploit their linguistic properties. We define the following generalized CI assumptions:

Assumption 1. *Given the most recent $n-1$ words and the classes of $m-1$ previous and k future words, a word W_t is independent of all previous words and their classes $\{W_1, \ldots, W_{t-n}, C_1, \ldots, C_{t-m}\}$.*

Assumption 2. *Given the most recent $n-1$ words and the class labels of previous $m-1$ words $\{C_{t-1}, \ldots, C_{t-m+1}, W_{t-1}, \ldots, W_{t-n+1}\}$, the class C_t is independent of previous words and distant class history $\{W_1, \ldots, W_{t-n}, C_1, \ldots, C_{t-m}\}$.*

The first assumption specifies the context of a word from both word and class variables allowing also the incorporation of the classes of future words. Schematically the



(a) Ass. 1 (b) Ass. 2

Fig. 2. Allowed dependencies due to assumption 1 and 2

allowed dependences are shown in Fig.2.a. The second assumption generalizes the class context to include word history. The schematic representation is shown in Fig.2.b.

The JPD for a specific model is: $P(W,C) = \prod_t P(W_t|\Pi_{W_t})P(C_t|\Pi_{C_t})$, where Π_{W_t} and Π_{C_t} are the set of parents of W_t and C_t respectively.

4 Experiments

The first set of experiments is performed on the *Le monde* newspaper corpus. We use 22M words for training and a test corpus of 2M words. The vocabulary consists of the most frequent 5000 words. The training corpus has been labeled automatically by a set of 200 syntactic classes set by hand [7]. All models used in the experiments are smoothed using absolute discounting method [8].

Table 1 shows perplexity results of different Bayesian network language models. In order to achieve our objective to find the best model we set the bi-class model as baseline and extend it incrementally by incorporating additional lexical and/or syntactic context. We also introduce the concept of right context of a word. DBN6 is a typical example of this case that integrates not only the left class context of a word but also its right syntactic context. We obtain a 16.6% improvement with respect to DBN4 that proves the importance of right context. It is true that linguistically this is not a surprising result. DBN5, on the other hand, shows that left context is quite important. That is why its removal reduces the results by 7.8%. A significant perplexity reduction is observed if a word not only depends on its syntactic but also lexical context. Indeed, DBN7 yields an improvement of 24.6% with respect to DBN4. This results confirms that lexical history is indispensable and that syntactic history provides a significant improvement.

Pushing forward this strategy, we achieve a model that is not only much better than the bi-class but also better than the bi-gram. Indeed, the model DBN8 reduces the perplexity by 57.9% with respect to bi-class and 2.4% with respect to bi-gram.

The second set of experiments (SPORT) is performed on articles dedicated to sport news extracted from a French newspaper. The corpus consists of 8500 sentences, the tests are performed using an open vocabulary of 2000 words with different cluster sizes ($|C|$)and with or without $< UNK >$. Word classes are defined based on statistical criteria using the HTK toolkit.

Table 1. Perplexity results on "Le Monde 87" corpus

Model	$P(W)$	PP
2-gram	$\prod_t P(w_t\|w_{t-1})$	65.24
2-class	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1})$	151.31
3-class	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1}c_{t-2})$	130.00
DBN4	$\prod_t P(w_t\|c_t,c_{t-1})P(c_t\|c_{t-1})$	113.13
DBN5	$\prod_t P(w_t\|c_t,c_{t+1})P(c_t\|c_{t-1})$	121.98
DBN6	$\prod_t P(w_t\|c_{t-1},c_t,c_{t+1})P(c_t\|c_{t-1})$	94.35
DBN7	$\prod_t P(w_t\|w_{t-1},c_t)P(c_t\|c_{t-1})$	85.20
DBN8	$\prod_t P(w_t\|w_{t-1},c_{t-1},c_{t-2})P(c_t\|w_t)$	**63.67**

Table 2. SPORT corpus perplexity results computed with/without $< UNK >$

Model	$P(W)$	$\|C = 1\|$	$\|C\| = 15$	$\|C\| = 25$	$\|C\| = 45$
2-gram	$\prod_t P(w_t\|w_{t-1})$	38.5/50.7			
3-gram	$\prod_t P(w_t\|w_{t-1}, w_{t-2})$	30.9/39.3			
2-class	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1})$		94.1/139.2	84.2/122.7	77.5/111.9
3-class	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1}, c_{t-2})$		90.6/133.4	80.0/115.9	71.5/102.2
DBN9	$\prod_t P(w_t\|w_{t-1}, c_{t-2})P(c_t\|w_t)$		36.3/47.2	35.2/45.6	34.4/44.3

Table 2 shows different model performances for SPORT corpus. The first remark is that the use of a higher number of classes leads to a reduction of perplexity. The second one is that the use of a history which combines classes and words is beneficial to language models and yields better results. The best performance is obtained by DBN9 which yields an improvement of 12,6% in comparison to bigram. The other important point is that, even if the trigram computational complexity ($O(|W|^3)$) is higher than the one of DBN9 ($O(|W|^2|C| + |W||C|)$), there is only a difference of 5 points (in average) between their perplexities, which is relatively small. Thus we can hope that, with a larger vocabulary and with a classification containing more classes, we can build DBN models similar to DBN9 with equivalent performances as a trigram.

5 Conclusion

Using the framework of the dynamic Bayesian networks, we presented a new approach for language modeling that considers data (training corpora made up of words, classes, concepts...) as observations of a dynamical system with the goal to find the model that has the best description of these observations in terms of perplexity. Among the advantages of this approach, we can note that the linguistic units are not used separately as in classical models, but merged in a single process. We tested several DBNs on different corpora and hence on different applications. The results show for all corpora that the models are improved by introducing the left context of both words and classes. Some experiments showed that DBNs outperform the baseline models and in some cases they compete with the higher order baseline models. All these encouraging results illustrate the feasibility of our approach. The main direction of our future work is to investigate algorithms of structure learning problem in order to reach our final objective: find the graph structure and its numerical parametrization that explains the data at best.

References

1. Jelinek, F. In: Self-organized language modeling for speech recognition. Morgan Kaufmann (1989) 450–506
2. Brown, P., DellaPietra, V., deSouza, P., Lai, J., Mercer, R.: Class based n-gram models of natural language. Computational Linguistics **18** (1992) 467–478
3. Heckerman, D.: A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division (1995)

4. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: UAI'98, Madison, Wisconsin (1998)
5. Deviren, M., Daoudi, K.: Structural learning of dynamic Bayesian networks in speech recognition. In: Eurospeech 2001, Aalborg, Denmark (2001)
6. Rosenfeld, R.: Adaptive Statistical Language Modeling: A Maximum Entropy Approach. PhD thesis, Carnegie Mellon University, Pittsburgh, PA 15213 (1994)
7. Smaïli, K., Brun, A., Zitouni, I., Haton, J.: Automatic and manual clustering for large vocabulary speech re cognition: A comparative study. In: Eurospeech, Hungary (1999)
8. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modelling. Computer Speech and Language **8** (1994) 1–38